Journal of Biomedical Informatics 47 (2014) 178–191

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Structural network analysis of biological networks for assessment of potential disease model organisms

Ahmed Ragab Nabhan [a,c,d], Indra Neil Sarkar [a,b,c,*]

[a] Center for Clinical & Translational Science, University of Vermont, Burlington, VT, USA
[b] Department of Microbiology & Molecular Genetics, University of Vermont, Burlington, VT, USA
[c] Department of Computer Science, University of Vermont, Burlington, VT, USA
[d] Faculty of Computers & Information, Fayoum University, Al Fayoum, Egypt

ARTICLE INFO

Article history:
Received 13 January 2013
Accepted 21 October 2013
Available online 5 November 2013

Keywords:
Disease pathway mining
Translational bioinformatics
Structural pattern analysis
Interaction networks

ABSTRACT

Model organisms provide opportunities to design research experiments focused on disease-related processes (e.g., using genetically engineered populations that produce phenotypes of interest). For some diseases, there may be non-obvious model organisms that can help in the study of underlying disease factors. In this study, an approach is presented that leverages knowledge about human diseases and associated biological interactions networks to identify potential model organisms for a given disease category. The approach starts with the identification of functional and interaction patterns of diseases within genetic pathways. Next, these characteristic patterns are matched to interaction networks of candidate model organisms to identify similar subsystems that have characteristic patterns for diseases of interest. The quality of a candidate model organism is then determined by the degree to which the identified subsystems match genetic pathways from validated knowledge. The results of this study suggest that non-obvious model organisms may be identified through the proposed approach.

1. Introduction

Complex diseases stem from an interplay of genetic and environmental factors. At the genetic level, these diseases are often associated with the dysfunction of more than one gene. This necessitates the study of complex diseases at a systems level, which includes the modeling of cellular processes that underlie an observed disorder and may involve both sequential and simultaneous molecular interactions between many agents (e.g., genes and chemical compounds). This highlights the importance of curating molecular interaction networks (e.g., gene/protein interaction networks, metabolic networks, and genetic pathways). Data resources that catalogue these networks are increasing both in terms of the number and size of networks as well as their coverage of organisms. Environmental factors, on the other hand, complicate the study of human diseases, since it is difficult to create a controlled environment that enables scientists to study environmental effects on disease development. Hence, model organisms offer opportunities for detailed study of features associated with complex diseases, because these organisms may be genetically engineered to produce desired phenotypes (e.g., associated with a particular disease of interest) and can be studied more easily in a controlled environment.

Model organisms play a vital role in advancing knowledge about disease processes. The sophisticated genetics of human diseases makes it important to study model organisms to uncover underlying mechanisms of diseases. Model organisms may not necessarily be closely related to humans from an evolutionary perspective. For instance, yeast are regularly used to model disease states [1]. Comparison of different phenotypes that arise from a conserved set of genes can be important for exploring model organisms for specific human disorders or diseases [2,3]. Analysis of model organism microarray data may also help identify those that have disease-related genes differentially expressed [2].

The house mouse (*Mus musculus*) has been a typical model organism in the study of human disease processes [4], as well as complex traits and social behavior [5]. Mice have also been genetically engineered to provide models for studying cancer and immune diseases [6,7]. However, mice may not always be suitable for the study of all categories of disease. In a recent study of 'phenologs' (phenotypes that are equivalent across organisms), McGary et al. suggested a worm model (*Caenorhabditis elegans*) for breast cancer, a mouse model for autism, a plant model (*Arabidopsis thaliana*) for Waardenburg syndrome, and a yeast model (*Saccharomyces cerevisiae*) for angiogenesis disorders [3]. Thus, there may be many potential choices for a suitable model organism relative to the spectrum of phenomena associated with disease. An

* Corresponding author. Address: Center for Clinical and Translational Science, University of Vermont, 89 Beaumont Avenue, Given Courtyard, N309, Burlington, VT 05405, USA. Fax: +1 802 656 4589.
E-mail address: neil.sarkar@uvm.edu (I.N. Sarkar).

1532-0464/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.10.011

empirical approach may therefore facilitate the identification of organism(s) that might provide insights to human diseases.

Evaluation of candidate model organisms might be measured by the degree to which gene/protein interaction networks include pathways that are structurally and functionally similar to human disease-related biological processes. To this end, prediction of pathways in candidate model organisms that are similar to disease-related pathways in humans can be effective in evaluating model organisms. Pathway prediction can be performed by a variety of techniques. A widely used technique involves mining gene or protein interaction networks to extract dense subgraphs (highly connected components within the network) and then calculating the statistical significance of the discovered subgraphs [8]. Statistically significant subgraphs are then cast as predicted pathways. Tian et al. developed a method to discover statistically significant pathways from gene expression data [9]. Bebek and Yang annotated gene networks with GO annotations and developed the Path-Finder method to predict novel pathways [10]. Cakmak and Ozsoyoglu developed a method that used frequent functional patterns in a known pathway to find organism-specific versions of that pathway in the gene networks [11]. Finally, Senf and Chen developed a hidden Markov model-based method to identify genes participating in genetic pathways [12].

The present study proposes a computational method that attempts to provide a quantitative measure of how well a candidate model organism might be suited for the study of a given disease type. The proposed quantitative measure is based on the proportion of correctly predicted genetic pathways that can be identified in interaction networks for a given organism. The proposed approach makes use of three types of knowledge resources: (1) Kyoto Encyclopedia of Gene and Genomes (KEGG) [13] pathway database, (2) The Biological General Repository for Interaction Datasets (BioGRID) and Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) gene/protein interaction databases [14], and (3) Gene Ontology (GO) [15] annotations that have been applied to genes or proteins in curated databases. The main premise of this work was to leverage a machine learning method to extract significant functional and structural patterns, or 'fingerprints,' [16] from functionally annotated KEGG disease pathways and match these patterns to functionally annotated gene/protein interaction networks in major databases (e.g., BioGRID) as well as meta-databases (e.g., STRING). Depending on an organism's interaction network coverage of structural patterns for a given disease, it can be ranked in terms of model organism suitability for that disease. Through the use of a statistical model, this study was able to quantify the dependency of functional structural patterns in pathways and disease categories for 14 organisms. It was assumed that some species may be a better suitable model for one disease category and thus less suitable for studying other diseases. This assumption was motivated by the McGary et al. study, where a range of model species were suggested for complex diseases [3]. The promising results suggest that the described approach may be used to determine the potential for a given organism to serve as a model for the study of a particular disease.

2. Materials and methods

In this section, the five phases of the developed approach are described: (1) annotation of gene/protein nodes in pathway graphs with molecular function annotations, (2) learning disease fingerprints within annotated pathways, (3) functional annotation and indexing of gene/protein interaction networks, (4) prediction of novel subsystems within gene/protein interaction networks using learned fingerprints, and (5) scoring discovered subsystems using reference pathways. Fig. 1 provides an overview of the approach.

2.1. Functional annotation of KEGG pathways

KEGG genetic pathways are modeled as directed graphs with a node set (V) representing biochemical entities such as genes, chemical compounds, and protein complexes and an edge set (E) representing interaction relations between entities such as general process type (e.g., a gene expression [GErel] or protein interaction [PPrel] relation) and specific relation types (e.g., activation, expression, and inhibition). For this study, only gene/protein nodes were considered. To increase the generalization capability, gene nodes were enriched with molecular function annotations as defined in Gene Ontology (GO) [15]. These GO annotations were imported from Human Protein Reference Database (HPRD) [17] and overlaid on gene/protein nodes of pathway graphs. Gene/protein nodes without a match to HPRD GO annotations were assigned a default 'NULL' annotation. Nodes could be associated with multiple GO term annotations and edges could also have multiple labels. Thus, for each graph there was a shift of focus from "what gene/protein is in a given node?" to "what function does the node perform in a system that models a biological process?" With knowledge-enriched annotations of genes/proteins, pathways were represented at a functional level. Subsequently, functional structural patterns in these pathways graphs could be matched to sub-networks of large interaction networks with functionally annotated nodes. In this study, the KEGG disease pathways dataset contained 63 disease pathways across seven human disease classes. KEGG disease pathways cover many biological processes related to genetic information processing, metabolism, and cellular processes. However, this study did not focus on a particular pathway category such as metabolic pathways and cellular processes. Each graph instance in this design set was associated with a class label from the seven disease classes in KEGG.

2.2. Learning disease fingerprints

The objective of the second module of the proposed method was to identify characteristic biological functionality patterns, termed "fingerprints," in annotated disease pathways. A mathematical model and an algorithm were designed to accomplish this task. A disease fingerprint was defined as a subgraph within a GO annotated disease pathway. Fingerprints were assumed to represent functional sub-processes that could be characteristic of a disease class such as immune, infectious, or neurodegenerative disease. Graphs in the design dataset were assumed to be independent and identically distributed (*iid*) data observed from an unknown probability distribution $P(G)$. The *iid* data assumption was made to facilitate statistical inference and to make decision about properties (e.g., class label) of a graph instance independent of other graph instances in the dataset. For a given GO-annotated pathway graph, there can be a large number of possible GO functionality subgraph patterns, which will be called "subgraph patterns" hereafter. A mathematical model was proposed to allow for scoring of subgraph patterns. High scoring patterns were output from the model as disease fingerprints.

Mining of key subgraph patterns in the dataset was performed so that a subgraph pattern is evaluated within a context of its neighboring patterns in a graph. To formalize the idea of neighbor context, a utility function termed "graph partitioning function" was used to decompose a graph into a set of subgraphs. A partitioning function π: $E(G) \rightarrow Z$ assigned an integer to every edge e of graph edge set $E(G)$ such that edges with the same integer formed a subgraph. The set of subgraphs $H\pi$ that were highlighted by a specific partitioning function (π) was defined as $H\pi = \{g_i | \forall e \in E(g_i),\ E(g_i) \subseteq E(G),\ \pi(e) = i\}$. Fig. 2 illustrates the concept of partitioning.